EXPONENTIAL QUANTUM ADVANTAGE IN COMMUNICATION FOR DISTRIBUTED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Training and inference with large machine learning models that far exceed the memory capacity of individual devices necessitates the design of distributed architectures, forcing one to contend with communication constraints. We present a framework for distributed computation over a quantum network in which data is encoded into specialized quantum states. We prove that for certain models within this framework, inference and training using gradient descent can be performed with exponentially less communication compared to their classical analogs, and with relatively modest time and space complexity overheads relative to standard gradient-based methods. To our knowledge, this is the first example of exponential quantum advantage for a generic class of machine learning problems with dense classical data that holds regardless of the data encoding cost. Moreover, we show that models in this class can encode highly nonlinear features of their inputs, and their expressivity increases exponentially with model depth. We also find that, interestingly, the communication advantage nearly vanishes for simpler linear classifiers. These results can be combined with natural privacy advantages in the communicated quantum states that limit the amount of information that can be extracted from them about the data and model parameters. Taken as a whole, these findings form a promising foundation for distributed machine learning over quantum networks.

1 INTRODUCTION

As the scale of the datasets and parameterized models used to perform computation over data continues to grow Kaplan et al. (2020); Hoffmann et al. (2022), distributing workloads across multiple devices becomes essential for enabling progress. The choice of architecture for large-scale training and inference must not only make the best use of computational and memory resources, but also contend with the fact that communication may become a bottleneck Pope et al. (2022). When using modern optical interconnects, classical computers communicate by sending bits represented by optical light. This however does not fully utilize the potential of the physical substrate as a communication resource. Given suitable computational capabilities and algorithms, the quantum nature of light can be harnessed as a powerful resource. Here we show that for a broad class of parameterized models, if quantum bits (*qubits*) are communicated instead of classical bits, an exponential reduction in the communication required to perform inference and gradient-based training can be achieved. This protocol additionally guarantees improved privacy of both the user data and model parameters through natural features of quantum mechanics, without the need for additional cryptographic or privacy protocols. To our knowledge, this is the first example of generic, exponential quantum advantage on problems that occur naturally in the training and deployment of large machine learning models. These types of communication advantages help scope the future roles and interplay between quantum and classical communication for distributed machine learning.

Quantum computers promise dramatic speedups across a number of computational tasks, with perhaps the most prominent example being the ability revolutionize our understanding of nature by enabling the simulation of quantum systems, owing to the natural similarity between quantum computers and the world Feynman (1982); Lloyd (1996). However, much of the data that one would like to compute with in practice seems to come from an emergent classical world rather than directly exhibiting quantum properties. While there are some well-known examples of exponential quantum speedups focusing on classical problems, most famously factoring Shor (1994) and related hidden subgroup problems Childs & van Dam (2008), these tend to be isolated and at times difficult to relate to practical applications. For example, even though significant speedups are known for certain ubiquitous problems in machine learning such as matrix inversion Harrow et al. (2009) and principal component analysis Lloyd et al. (2014), the advantage is often lost when including the cost of loading classical data into the quantum computer or of reading out the result into classical

memory Aaronson (2015). In applications where an efficient data access model avoids the above pitfalls, the complexity of quantum algorithms tends to depend on condition numbers of matrices which scale with system size in a way that reduces or even eliminates any quantum advantage Montanaro & Pallister (2015). It is worth noting that much of the discussion about the impact of quantum technology on machine learning has focused on computational advantage. However quantum resources are not only useful in reducing computational complexity — they can also provide an advantage in communication complexity, enabling exponential reductions in communication for some problems Raz (1999); Bar-Yossef et al. (2008). Inspired by these results, we study a setting where quantum advantage in communication is possible across a wide class of machine learning models. This advantage holds without requiring any sparsity assumptions or elaborate data access models such as QRAM Giovannetti et al. (2008).

We focus on compositional distributed learning, known as *pipelining* Huang et al. (2018); Barham et al. (2022). While there are a number of strategies for distributing machine learning workloads that are influenced by the requirements of different applications and hardware constraints Xu et al. (2021); Jouppi et al. (2023), splitting up a computational graph in a compositional fashion (Figure 1) is a common approach. We describe distributed, parameterized quantum circuits that can be used to perform inference over data when distributed in this way, and can be trained using gradient methods (Section 2.1). The ideas we present can also be used to optimize models that use data parallelism as well. In principle, such circuits could be implemented on quantum computers that are able to communicate quantum states.

For the general class of distributed, parameterized quantum circuits that we study, we show the following:

- Even for simple circuits in this class, there is an exponential quantum advantage in communication for the problem of estimating the loss and the gradients of the loss with respect to the parameters (Section 3). This additionally implies a privacy advantage from Holevo's bound (Section 6).
- For a subclass of these circuits, there is an exponential advantage in communication for the entire training process, not just gradient estimation. This subclass includes circuits for fine-tuning using pre-trained features, and the proof is constructed by through convergence rates under convexity assumptions (Section 4).
- The ability to interleave multiple unitaries encoding nonlinear features of data enables expressivity to grow exponentially with depth, and universal function approximation in some settings implying these models are highly expressive in contrast to popular belief about linear restrictions in quantum neural networks (Section 5).

2 Preliminaries

2.1 Large-scale learning problems and distributed computation

Pipelining is a commonly used method of distributing a machine learning workload, in which different layers of a deep model are allocated distinct hardware resources Huang et al. (2018); Narayanan et al. (2019). Training and inference then require communication of features between nodes. Pipelining enables flexible changes to the model architecture in a task-dependent manner, since subsets of a large model can be combined in an adaptive fashion to solve many downstream tasks. Additionally, pipelining allows sparse activation of a subset of a model required to solve a task, and facilitates better use of heterogeneous compute resources since it does not require storing identical copies of a large model. The potential for large models to be easily fine-tuned to solve multiple tasks is well-known Brown et al. (2020); Bommasani et al. (2021), and pipelined architectures which facilitate this are the norm in the latest generation of large language models Rasley et al. (2020); Barham et al. (2022). Data parallelism, in contrast, involves storing multiple copies of the model on different nodes, training each on a subsets of the data and exchanging information to synchronize parameter updates. In practice, different parallelization strategies are combined in order to exploit trade-offs between latency and throughput in a task-dependent fashion Xu et al. (2021); Jouppi et al. (2023); Pope et al. (2022).

2.2 Communication complexity

Communication complexity Yao (1979); Kushilevitz & Nisan (2011); Rao & Yehudayoff (2020) is the study of distributed computational problems using a cost model that focuses on the communication required between players rather than the time or computational complexity. It is naturally related to the study of the space complexity of streaming algorithms Roughgarden (2015). The key object of study in this area is the tree induced by a communication protocol whose nodes enumerate all possible communication histories and whose leaves correspond to the outputs of

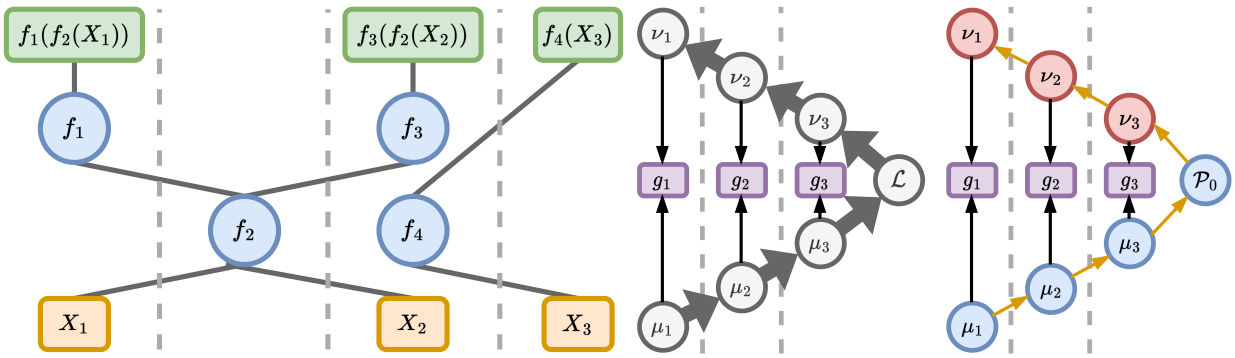

Figure 1: *Left:* Distributed, compositional computation. Dashed lines separate devices with computational and storage resources. The circular nodes represent parameterized functions that are allocated distinct hardware resources and are spatially separated, while the square nodes represent data (yellow) and outputs corresponding to different tasks (green). The vertical axis represents time. This framework of hardware allocation enables flexible modification of the model structure in a task-dependent fashion. *Right:* Computation of gradient estimators g_ℓ at different layers of a model distributed across multiple devices by pipelining. Computing forward features μ_ℓ and backwards features ν_ℓ (also known as computing a forward or backward pass) requires a large amount of classical communication (grey) but an exponentially smaller amount of quantum communication (yellow). $\mathcal{L}$ is the classical loss function, and $\mathcal{P}_0$ an operator whose expectation value with respect to a quantum model gives the analogous loss function in the quantum case.

the protocol. The product structure induced on the leaves of this tree as a function of the inputs allows one to bound the depth of the tree from below, which gives an unconditional lower bound on the communication complexity. The power of replacing classical bits of communication with qubits has been the subject of extensive study Chi-Chih Yao (1993); Brassard (2001); Buhrman et al. (2009). For certain problems such as Hidden Matching Bar-Yossef et al. (2008) and a variant of classification with deep linear models Raz (1999) an exponential quantum communication advantage holds, while for other canonical problems such as Disjointness only a polynomial advantage is possible Razborov (2002).

At a glance, the development of networked quantum computers may seem much more challenging than the already herculean task of building a fault tolerant quantum computer. However, for some quantum network architectures, the existence of a long-lasting fault tolerant quantum memory as a quantum repeater, may be the enabling component that lifts low rate shared entanglement to a fully functional quantum network Munro et al. (2015), and hence the timelines for small fault tolerant quantum computers and quantum networks may be more coincident than widely believed. As such, it is well motivated to consider potential communication advantages alongside computational advantages when talking about the applications of fault tolerant quantum computers. In Appendix G we briefly survey approaches to implementing quantum communication in practice, and the associated challenges.

In addition, while we largely restrict ourselves here to discussions of communication advantages, and most other studies focus on purely computational advantages, there may be interesting advantages at their intersection. For example, it is known that no quantum state built from a simple (or polynomial complexity) circuit can confer an exponential communication advantage, however states made from simple circuits can be made computationally difficult to distinguish Ji et al. (2018). Hence the use of quantum pre-computation Huggins & McClean (2023) and communication may confer advantages even when traditional computational and communication cost models do not admit such advantages do to their restriction in scope.

3 Distributed Learning with Quantum Resources

In this work we focus on parameterized models that are representative of the most common models used and studied today in quantum machine learning, sometimes referred to as quantum neural networks McClean et al. (2015); Farhi & Neven (2018); Cerezo et al. (2020); Schuld et al. (2020). We will use the standard Dirac notation of quantum mechanics throughout. A summary of relevant notation and the fundamentals of quantum mechanics is provided in Appendix A. We define a class models with parameters Θ, taking an input x which is a tensor of size N. The models take the following general form:

Definition 3.1. *$\{A_\ell(\theta_\ell^A, x)\}, \{B_\ell(\theta_\ell^B, x)\}$ for $\ell \in \{1, \ldots, L\}$ are each a set of unitary matrices of size $N' \times N'$ for some N' such that $\log N' = O(\log N)$*[1]*. The $\theta_\ell^A, \theta_\ell^B$ are vectors of P parameters each. For every ℓ, i, we assume that $\frac{\partial A_\ell}{\partial \theta_{\ell i}^A}$ is anti-hermitian up to a real scaling factor and has at most two eigenvalues, and similarly for B_ℓ.*

The model we consider is defined by

$$|\varphi(\Theta, x)\rangle \equiv \left(\prod_{\ell=L}^{1} A_\ell(\theta_\ell^A, x) B_\ell(\theta_\ell^B, x) \right) |\psi(x)\rangle \,, \tag{3.1}$$

where $\psi(x)$ is a fixed state of $\log N'$ qubits.

The loss function is given by

$$\mathcal{L}(\Theta, x) \equiv \langle \varphi(\Theta, x)| \, \mathcal{P}_0 \, |\varphi(\Theta, x)\rangle \,, \tag{3.2}$$

where $\mathcal{P}_0$ is a Pauli matrix that acts on the first qubit.

In standard linear algebra notation, the output of the model is a unit norm N'-dimensional complex vector φ_L, defined recursively by

$$\varphi_0 = \psi(x), \quad \varphi_\ell = A_\ell(\theta_\ell^A, x) B_\ell(\theta_\ell^B, x) \varphi_{\ell-1}, \tag{3.3}$$

where the entries of φ_L are represented by the amplitudes of a quantum state. The loss takes the form $\mathcal{L}(\Theta, x) = (\varphi_L^*)^T \mathcal{P}_0 \varphi_L$, which includes the standard L^2 loss as a special case.

Subsequently we omit the dependence on x and Θ (or subsets of it) to lighten notation, and consider special cases where only subsets of the unitaries depend on x, or where the unitaries take a particular form and may not be parameterized. Denote by $\nabla_{A(B)} \mathcal{L}$ the entries of the gradient vector that correspond to the parameters of $\{A_\ell\}(\{B_\ell\})$.

While seemingly stringent, the condition on the derivatives is in fact satisfied by many of the most common quantum neural network architectures Cerezo et al. (2020); Crooks (2019); Schuld et al. (2020). This condition is satisfied for example if

$$A_\ell = \prod_{j=1}^{P} e^{i \alpha_{\ell j}^A \theta_{\ell j}^A \mathcal{P}_{\ell j}^A} \tag{3.4}$$

and the $\mathcal{P}_{\ell j}^A$ are both unitary and Hermitian (e.g. Pauli matrices), while $\beta_{\ell j}^A$ are scalars. Such models, or parameterized quantum circuits, are naturally amenable to implementation on quantum devices , and for $P = O(N^2)$ any unitary over $\log N'$ qubits can be written in this form. In the special case where x in a unit norm N-dimensional vector, a simple choice of $|\psi(x)\rangle$ is the amplitude encoding of x, given by

$$|\psi(x)\rangle = |x\rangle = \sum_{i=0}^{N-1} x_i \, |i\rangle \,. \tag{3.5}$$

However, despite its exponential compactness in representing the data, a naive implementation of the simplest choice is restricted to representing quadratic features of the data that can offer no substantial quantum advantage in a learning task Huang et al. (2021), so the choice of data encoding is critical to the power of a model. The interesting parameter regime for classical data and models is one where N, P are large, while L is relatively modest. For general unitaries $P = O(N^2)$, which matches the scaling of the number of parameters in fully-connected networks. When the input tensor x is a batch of datapoints, N is equivalent to the product of batch size and input dimension.

The model in Definition 3.1 can be used to define distributed inference and learning problems by dividing the input x and the parameterized unitaries between two players, Alice and Bob. We define their respective inputs as follows:

$$\begin{aligned} &\text{Alice}: \quad |\psi(x)\rangle \,, \{A_\ell\}, \\ &\text{Bob}: \quad \{B_\ell\}. \end{aligned} \tag{3.6}$$

The problems of interest require that Alice and Bob compute certain joint functions of their inputs. As a trivial base case, it is clear that in a communication cost model, all problems can be solved with communication cost at most the size of the inputs times the number of parties, by a protocol in which each party sends its inputs to all others. We will be interested in cases where one can do much better by taking advantage of quantum communication.

[1] We will consider some cases where $N' = N$, but will find it helpful at times to encode nonlinear features of x in these unitaries, in which case we may have $N' > N$.

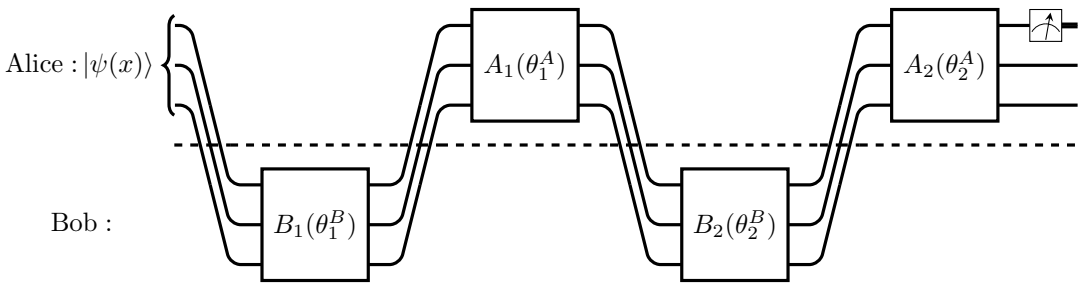

Figure 2: Distributed quantum circuit implementing $\mathcal{L}$ for $L = 2$. Both $\mathcal{L}$ and its gradients with respect to the parameters of the unitaries can be estimated with total communication that is logarithmic in the number of amplitudes N and the number of parameters per unitary P.

Given the inputs eq. (3.6), we will be interested chiefly in the two problems specified below.

Problem 1 (Distributed Inference)**.** *Alice and Bob each compute an estimate of $\langle\varphi| \mathcal{P}_0 |\varphi\rangle$ up to additive error ε.*

The straightforward algorithm for this problem, illustrated in fig. 2, requires L rounds of communication. The other problem we consider is the following:

Problem 2 (Distributed Gradient Estimation)**.** *Alice computes an estimate of $\nabla_A \langle\varphi| \mathcal{P}_0 |\varphi\rangle$, while Bob computes an estimate of $\nabla_B \langle\varphi| Z_0 |\varphi\rangle$, up to additive error ε in L^∞.*

3.1 Communication complexity of inference and gradient estimation

We show that inference and gradient estimation are achievable with a logarithmic amount of quantum communication, which will represent an exponential improvement over the classical cost for some cases:

Lemma 1. *Problem 1 can be solved by communicating $O(\log N)$ qubits over $O(L/\varepsilon^2)$ rounds.*

Proof: Appendix B.

Lemma 2. *Problem 2 can be solved with probability greater than $1 - \delta$ by communicating $\tilde{O}(\log N (\log P)^4 \log(L/\delta)/\varepsilon^4)$ qubits over $O(L^2)$ rounds. The time and space complexity of the algorithm is $\sqrt{P}\, L\, \mathrm{poly}(N, \log P, \varepsilon^{-1}, \log(1/\delta))$.*

Proof: Appendix B.

This upper bound is obtained by simply noting that the problem of gradient estimation at every layer can be reduced to a shadow tomography problem:

Theorem 1 (Shadow Tomography Aaronson (2017b))**.** *For an unknown state $|\psi\rangle$ of $\log N$ qubits, given K known two-outcome measurements E_i, there is an explicit algorithm that takes $|\psi\rangle^{\otimes k}$ as input, where $k = \tilde{O}(\log^4 K \log N \log(1/\delta)/\varepsilon^4)$, and produces estimates of $\langle\psi| E_i |\psi\rangle$ for all i up to additive error ε with probability greater than $1 - \delta$. $\tilde{O}$ hides a factor $\mathrm{poly}(\log\log K, \log\log N, \log 1/\varepsilon)$.*

Using immediate reductions from known problems in communication complexity, we can show that the amount of classical communication required to solve these problem is polynomial in the size of the input, and additionally give a lower bound on the number of rounds of communication required by any quantum or classical algorithm:

Lemma 3. *i) The classical communication complexity of Problem 1 and Problem 2 is $\Omega(\max(\sqrt{N}, L))$.*

ii) Any algorithm (quantum or classical) for Problem 1 or Problem 2 requires either $\Omega(L)$ rounds of communication or $\Omega(N/L^4)$ qubits (or bits) of communication.

Proof: Appendix B.

The implication of the second result in Lemma 3 is that $\Omega(L)$ rounds of communication are necessary in order to obtain an exponential communication advantage for small L, since otherwise the number of qubits of communication required can scale linearly with N.

Combining Lemma 2 and Lemma 3, in the regime where $L = O(\text{polylog}(N))$, which is relevant for classical machine learning models, we obtain an exponential advantage in communication complexity for both inference and gradient estimation. The required overhead in terms of time and space is only polynomial when compared to the straightforward classical algorithms for these problems.

The distribution of the model as in eq. (3.6) is an example of pipelining. Data parallelism is another common approach to distributed machine learning in which subsets of the data are distributed to identical copies of the model. In Appendix C we show that it can also be implemented using quantum circuits, which can then trained using gradient descent requiring quantum communication that is logarithmic in the number of parameters and input size.

Quantum advantage is possible in these problems because there is a bound on the complexity of the final output, whether it be correlated elements of the gradient up to some finite error or the low-dimensional output of a model. This might lead one to believe that whenever the output takes such a form, encoding the data in the amplitudes of a quantum state will trivially give an exponential advantage in communication complexity. We show however that the situation is slightly more nuanced, by considering the problem of inference with a linear model:

Lemma 4. *For the problem of distributed linear classification, there can be no exponential advantage in using quantum communication in place of classical communication.*

The precise statement and proof of this result are presented in Appendix E. This result also highlights that the worst case lower bounds such as Lemma 3 may not hold for circuits with certain low-dimensional or other simplifying structure.

4 Exponential advantages in end-to-end training

So far we have discussed the problems of inference and estimating a single gradient vector. It is natural to also consider when these or other gradient estimators can be used to efficiently solve an optimization problem (i.e. when the entire training processes is considered rather than a single iteration). Applying the gradient estimation algorithm detailed in Lemma 2 iteratively gives a distributed stochastic gradient descent algorithm which we detail in Algorithm 2, yet one may be concerned that a choice of $\varepsilon = O(\log N)$ which is needed to obtain an advantage in communication complexity will preclude efficient convergence. Here we present a simpler algorithm that requires a single quantum measurement per iteration, and can provably solve certain convex problems efficiently, as well as an application of shadow tomography to fine-tuning where convergence can be guaranteed, again with only logarithmic communication cost. In both cases, there is an exponential advantage in communication even when considering the entire training process.

4.1 "Smooth" circuits

Consider the case where A_ℓ are product of rotations for all ℓ, namely

$$A_\ell = \prod_{j=1}^{P} e^{-\frac{1}{2} i \beta_{\ell j}^A \theta_{\ell j}^A \mathcal{P}_{\ell j}^A}, \tag{4.1}$$

where $\mathcal{P}_{\ell j}^A$ are Pauli matrices acting on all qubits, and similarly for B_ℓ. These can also be interspersed with other non-trainable unitaries. This constitutes a slight generalization of the setting considered in Harrow & Napp (2021), and the algorithm we present is essentially a distributed distributed version of theirs. Denote by β an $2PL$-dimensional vector with elements $\beta_{\ell j}^Q$ where $Q \in \{A, B\}$ [2]. The quantity $\|\beta\|_1$ is the total evolution time if we interpret the state $|\varphi\rangle$ as a sequence of Hamiltonians applied to the initial state $|x\rangle$.

In Appendix D.1 we describe an algorithm that converges to the neighborhood of a minimum, or achieves $\mathbb{E}\mathcal{L}(\Theta) - \mathcal{L}(\Theta^\star) \leq \varepsilon_0$, for a convex $\mathcal{L}$ after

$$\frac{2 \left\|\Theta^{(0)} - \Theta^\star\right\|_2^2 \|\beta\|_1^2}{\varepsilon_0^2} \tag{4.2}$$

[2] Harrow & Napp (2021) actually consider a related quantity for which has smaller norm in cases where multiple gradient measurements commute, leading to even better rates.

iterations, where $\Theta^\star$ are the parameter values at the minimum of $\mathcal{L}$. The expectation is with respect to the randomness of quantum measurement and additional internal randomness of the algorithm. The algorithm is based on classically sampling a single coordinate to update at every iteration, and computing an unbiased estimator of the gradient with a single measurement. It can thus be seen as a form of probabilistic coordinate descent.

This implies an exponential advantage in communication for the entire training process as long as $\left\|\Theta^{(0)}-\Theta^\star\right\|_2^2\|\beta\|_1^2 = \mathrm{polylog}(N)$. Such circuits either have a small number of trainable parameters ($P = O(\mathrm{polylog}(N))$), else depend weakly on each parameter (e.g. $\beta_{\ell j}^Q = O(1/P)$ for arbitrary P), or have structure that allows initial parameter guesses whose quality diminishes quite slowly with system size. Nevertheless, over a convex region the loss can rapidly change by an $O(1)$ amount. One may also be concerned that in the setting $\left\|\Theta^{(0)}-\Theta^\star\right\|_2^2\|\beta\|_1^2 = \mathrm{polylog}(N)$ only a logarithmic number of parameters is updated during the entire training process and so the total effect of the training process may be negligible. It is important to note however that each such sparse update depends on the structure of the entire gradient vector as seen in the sampling step. In this sense the algorithm is a form of probabilistic coordinate descent, since the probability of updating a coordinate $|\beta_{\ell j}^Q|/\left\|\beta\right\|_1$ is proportional to the the magnitude of the corresponding element in the gradient (actually serving as an upper bound for it).

Remarkably, the time complexity of a single iteration of this algorithm is proportional to a forward pass, and so matches the scaling of classical backpropagation. This is in contrast to the polynomial overhead of shadow tomography (Theorem 1). Additionally, it requires a single measurement per iteration, without any of the additional factors in the sample complexity of shadow tomography.

4.2 Fine-tuning the last layer of a model

Consider a model given by eq. (3.1) where only the parameters of A_L are trained, and the rest are frozen, and denote this model by $|\varphi_f\rangle$. The circuit up to that unitary could include multiple data-dependent unitaries that represent complex features in the data. Training only the final layer in this manner is a common method of fine-tuning a pre-trained model Howard & Ruder (2018). If we now define

$$E_{Li}^A = |1\rangle\langle 0|\otimes A_L^\dagger \mathcal{P}_0 \frac{\partial A_L}{\partial\theta_{Li}^A} + |0\rangle\langle 1|\otimes\left(\frac{\partial A_L}{\partial\theta_{Li}^A}\right)^\dagger \mathcal{P}_0 A_L, \tag{4.3}$$

the expectation value of E_{Li}^A using the state $|+\rangle\left|\mu^{A_L}\right\rangle$ gives $\frac{\partial\mathcal{L}}{\partial\theta_{\ell i}^A}$. Here

$$\left|\mu_L^A\right\rangle = B_L(x)\prod_{k=L-1}^{1} A_k(x)B_k(x)\,|\psi(x)\rangle \tag{4.4}$$

is the forward feature computed by Alice at layer L with the parameters of all the other unitaries frozen (hence the dependence on them is dropped). Since the observables in the shadow tomography problem can be chosen in an online fashion, and adaptively based on previous measurements, we can simply define a stream of measurement operators by measuring P observables to estimate the gradients w.r.t. an initial set of parameters, updating these parameters using gradient descent with step size η, and defining a new set of observables using the updated parameters. Repeating this for T iterations gives a total of PT observables. By the scaling in Lemma 2, the total communication needed is $\tilde{O}(\log N(\log TP)^4\log(L/\delta)/\varepsilon^4)$ over $O(L)$ rounds (since only $O(L)$ rounds are needed to create copies of $\left|\mu^{A_L}\right\rangle$. This implies an exponential advantage in communication for the entire training process (under the reasonable assumption $T = O(\mathrm{poly}(N,P))$), despite the additional stochasticity introduced by the need to perform quantum measurements. For example, assume one has a bound $\|\nabla\mathcal{L}\|_2^2 \le K$. If the circuit is comprised of unitaries with Hermitian derivatives, this holds with $K = PL$. In that case, denoting by g the gradient estimator obtained by shadow tomography, we have

$$\|g\|_2^2 \le \|\nabla\mathcal{L}\|_2^2 + \|\nabla\mathcal{L}-g\|_2^2 \le K + \varepsilon^2 PL. \tag{4.5}$$

It then follows directly from Lemma 6 that for an appropriately chosen step size, if $\mathcal{L}$ is convex one can find parameter values $\overline{\Theta}$ such that $\mathcal{L}(\overline{\Theta}) - \mathcal{L}(\Theta^\star) \le \varepsilon_0$ using

$$T = 2\frac{\left\|\Theta^{(0)}-\Theta^\star\right\|_2^2 (K+\varepsilon^2 PL)^2}{\varepsilon_0^2} \tag{4.6}$$

iterations of gradient descent. Similarly if $\mathcal{L}$ is λ-strongly convex then $T = 2(K + \varepsilon^2 PL)^2/\lambda\varepsilon_0 + 1$ iterations are sufficient. In both cases therefore an exponential advantage is achieved for the optimization process as a whole, since in both cases one can implement the circuit that is used to obtain the lower bounds in Lemma 3.

5 Expressivity of compositional models

It is natural to ask how expressive models of the form of eq. (3.1) can be, given the unitarity constraint of quantum mechanics on the matrices $\{A_\ell, B_\ell\}$. This is a nuanced question that can depend on the encoding of the data that is chosen and the method of readout. On the one hand, if we pick $|\psi(x)\rangle$ as in eq. (3.5) and use $\{A_\ell, B_\ell\}$ that are independent of x, the resulting state $|\varphi\rangle$ will be a linear function of x. On the other hand, one could map bits to qubits 1-to-1 and encode any reversible classical function of data within the unitary matrices $\{A_\ell(x)\}$ with the use of extra space qubits. However, this negates the possibility of any space or communication advantages (and does not provide any real computational advantage without additional processing). As above, one prefers to work on more generic functions in the amplitude and phase space, allowing for an exponential compression of the data into a quantum state, but one that must be carefully worked with.

We investigate the consequences of picking $\{A_\ell(x)\}$ that are *nonlinear* functions of x, and $\{B_\ell\}$ that are data-independent. This is inspired by a common use case in which Alice holds some data or features of the data, while Bob holds a model that can process these features. Given a scalar variable x, define $A_\ell(x) = \mathrm{diag}((e^{-2\pi i\lambda_{\ell 1}x}, \ldots, e^{-2\pi i\lambda_{\ell N}x}))$ for $\ell \in \{1, \ldots, L\}$. We also consider parameterized unitaries $\{B_\ell\}$ that are independent of the $\{\lambda_{\ell i}\}$ and inputs x, y, and the state obtained by interleaving the two in the manner of eq. (3.1) by $|\varphi(x)\rangle$.

We next set $\lambda_{\ell 1} = 0$ for all $\ell \in \{1, \ldots, L\}$ and $\lambda_{L2} = 0$. If we are interested in expressing the frequency

$$\Lambda_{\bar{j}} = \sum_{\ell=1}^{L-1} \lambda_{\ell j_\ell}, \tag{5.1}$$

where $j_\ell \in \{2, \ldots, N\}$, we simply initialize with $|\psi(x)\rangle = |+\rangle_0 |0\rangle$ and use

$$B_\ell = |j_\ell - 1\rangle \langle j_{\ell-1} - 1| + |j_{\ell-1} - 1\rangle \langle j_\ell - 1|, \tag{5.2}$$

with $j_1 = j_L = 2$. It is easy to check that the resulting state is $|\varphi(x)\rangle = \left(|0\rangle + e^{-2\pi i\Lambda_{\bar{j}}x} |1\rangle\right)/\sqrt{2}$. Since the basis state $|0\rangle$ does not accumulate any phase, while the B_ℓs swap the $|1\rangle$ state with the appropriate basis state at every layer in order to accumulate a phase corresponding to a single summand in eq. (5.1). Choosing to measure the operator $\mathcal{P}_0 = X_0$, it follows that $\langle\varphi(x)| X_0 |\varphi(x)\rangle = \cos(2\pi\Lambda_{\bar{j}}x)$.

It is possible to express $O(N^L)$ different frequencies in this way, assuming the $\Lambda_{\bar{j}}$ are distinct, which will be the case for example with probability 1 if the $\{\lambda_{\ell i}\}$ are drawn i.i.d. from some distribution with continuous support. This further motivates the small L regime where exponential advantage in communication is possible. These types of circuits with interleaved data-dependent unitaries and parameterized unitaries was considered for example in Schuld et al. (2020), and is also related to the setting of quantum signal processing and related algorithms Low & Chuang (2017); Martyn et al. (2021). We also show that such circuits can express dense function in Fourier space, and for small N we additionally find that these circuits are universal function approximators (Appendix F.1), though in this setting the possible communication advantage is less clear.

The problem of applying nonlinearities to data encoded efficiently in quantum states is non-trivial and is of interest due to the importance of nonlinearities in enabling efficient function approximation Maiorov & Pinkus (1999). One approach to resolving the constraints of unitarity with the potential irreversibility of nonlinear functions is the introduction of slack variables via additional ancilla qubits, as typified by the techniques of block-encoding Chakraborty et al. (2018); Gilyén et al. (2018). Indeed, these techniques can be used to apply nonlinearities to amplitude encoded data efficiently, as was recently shown in Rattew & Rebentrost (2023). This approach can be applied to the distributed setting as well. Consider the communication problem where Alice is given x as input and Bob is given unitaries $\{U_1, U_2\}$ over $\log N$ qubits. Denote by $\sigma : \mathbb{R} \to \mathbb{R}$ a nonlinear function such as the sigmoid, exponential or standard trigonometric functions, and $n = 2^N$. We show the following:

Lemma 5. *There exists a model $|\varphi_\sigma\rangle$ of the form definition 3.1 with $L = O(\log 1/\varepsilon)$, $N' = 2^{n'}$ where $n' = 2n + 4$ such that*

$$|\varphi_\sigma\rangle = \alpha |0\rangle^{\otimes n+4} |\hat{y}\rangle + |\phi\rangle \tag{5.3}$$

for some $\alpha = O(1)$, where $|\hat{y}\rangle$ is a state that obeys

$$\left\| |\hat{y}\rangle - \left| U_2 \frac{1}{\|\sigma(U_1 x)\|_2} \sigma(U_1 x) \right\rangle \right\|_2 < \varepsilon. \tag{5.4}$$

$|\phi\rangle$ is a state whose first $n+4$ registers are orthogonal to $|0\rangle^{\otimes n+4}$.

Proof: Appendix B.

This result implies that with constant probability, after measurement of the first $n+4$ qubits of $|\varphi_\sigma\rangle$, one obtains a state whose amplitudes encode the output of a single hidden layer neural network.

It is also worth noting that the general form of the circuits we consider resembles self-attention based models with their nonlinearities removed (motivated for example by Sun et al. (2023)), as we explain in Appendix F.2. Finally, in Appendix F.3 we discuss other strategies for increasing the expressivity of these quantum circuits by combining them with classical networks.

6 Privacy of Quantum Communication

In addition to an advantage in communication complexity, the quantum algorithms outlined above have an inherent advantage in terms of privacy. It is well known that the number of bits of information that can be extracted from an unknown quantum state is proportional to the number of qubits. It follows immediately that since the above algorithm requires exchanging a logarithmic number of copies of states over $O(\log N)$ qubits, even if all the communication between the two players is intercepted, an attacker cannot extract more than a logarithmic number of bits of classical information about the input data or model parameters. Specifically, we have:

Corollary 1. *If Alice and Bob are implementing the quantum algorithm for gradient estimation described in Lemma 2, and all the communication between Alice and Bob is intercepted by an attacker, the attacker cannot extract more than $\tilde{O}(L^2(\log N)^2(\log P)^4 \log(L/\delta)/\varepsilon^4)$ bits of classical information about the inputs to the players.*

This follows directly from Holevo's theorem Holevo (1973), since the multiple copies exchanged in each round of the protocol can be thought of as a quantum state over $\tilde{O}((\log N)^2(\log P)^4 \log(L/\delta)/\varepsilon^4)$ qubits. As noted in Aaronson (2017b), this does not contradict the fact that the protocol allows one to estimate all P elements of the gradient, since if one were to place some distribution over the inputs, the induced distribution over the gradient elements will generally exhibit strong correlations. An analogous result holds for the inference problem described in Lemma 1.

It is also interesting to ask how much information either Bob or Alice can extract about the inputs of the other player by running the protocol. If this amount is logarithmic as well, it provides additional privacy to both the model owner and the data owner. It allows two actors who do not necessarily trust each other, or the channel through which they communicate, to cooperate in jointly training a distributed model or using one for inference while only exposing a vanishing fraction of the information they hold.

It is also worth mentioning that data privacy is also guaranteed in a scenario where the user holding the data also specifies the processing done on the data. In this setting, Alice holds both data x and a full description of the unitaries she wishes to apply to her state. She can send Bob a classical description of these unitaries, and as long as the data and features are communicated in the form of quantum states, only a logarithmic amount of information can be extracted about them. In this setting there is of course no advantage in communication complexity, since the classical description of the unitary will scale like $\mathrm{poly}(N, P)$.

7 Discussion

This work constitutes a preliminary investigation into a generic class of quantum circuits that has the potential for enabling an exponential communication advantage in problems of classical data processing including training and inference with large parameterized models over large datasets. Communication constraints may become even more relevant as models continue to grow, and in setting they are trained on data that is obtained by distributed interaction with the physical world Driess et al. (2023). Our results also naturally raise further questions regarding the expressive power and trainability of these types of circuits, which may be of independent interest. We collect some of these in Appendix H.

REFERENCES

Scott Aaronson. Read the fine print. *Nat. Phys.*, 11(4):291–293, April 2015.

Scott Aaronson. Introduction to quantum information science. https://www.scottaaronson.com/qclec.pdf, 2017a.

Scott Aaronson. Shadow tomography of quantum states. November 2017b.

Amira Abbas, Robbie King, Hsin-Yuan Huang, William J Huggins, Ramis Movassagh, Dar Gilboa, and Jarrod R McClean. On quantum backpropagation, information reuse, and cheating measurement collapse. May 2023.

Dimitris Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J. Comput. System Sci.*, 66(4):671–687, June 2003.

Alekh Agarwal, Peter L Bartlett, Pradeep Ravikumar, and Martin J Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. September 2010.

Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C Bardin, Rami Barends, Rupak Biswas, Sergio Boixo, Fernando G S L Brandao, David A Buell, Brian Burkett, Yu Chen, Zijun Chen, Ben Chiaro, Roberto Collins, William Courtney, Andrew Dunsworth, Edward Farhi, Brooks Foxen, Austin Fowler, Craig Gidney, Marissa Giustina, Rob Graff, Keith Guerin, Steve Habegger, Matthew P Harrigan, Michael J Hartmann, Alan Ho, Markus Hoffmann, Trent Huang, Travis S Humble, Sergei V Isakov, Evan Jeffrey, Zhang Jiang, Dvir Kafri, Kostyantyn Kechedzhi, Julian Kelly, Paul V Klimov, Sergey Knysh, Alexander Korotkov, Fedor Kostritsa, David Landhuis, Mike Lindmark, Erik Lucero, Dmitry Lyakh, Salvatore Mandrà, Jarrod R McClean, Matthew McEwen, Anthony Megrant, Xiao Mi, Kristel Michielsen, Masoud Mohseni, Josh Mutus, Ofer Naaman, Matthew Neeley, Charles Neill, Murphy Yuezhen Niu, Eric Ostby, Andre Petukhov, John C Platt, Chris Quintana, Eleanor G Rieffel, Pedram Roushan, Nicholas C Rubin, Daniel Sank, Kevin J Satzinger, Vadim Smelyanskiy, Kevin J Sung, Matthew D Trevithick, Amit Vainsencher, Benjamin Villalonga, Theodore White, Z Jamie Yao, Ping Yeh, Adam Zalcman, Hartmut Neven, and John M Martinis. Quantum supremacy using a programmable superconducting processor. *Nature*, 574 (7779):505–510, October 2019.

Koji Azuma, Sophia E Economou, David Elkouss, Paul Hilaire, Liang Jiang, Hoi-Kwong Lo, and Ilan Tzitrin. Quantum repeaters: From quantum networks to the quantum internet. December 2022.

Krishna C Balram and Kartik Srinivasan. Piezoelectric optomechanical approaches for efficient quantum microwave-to-optical signal transduction: the need for co-design. August 2021.

Ziv Bar-Yossef, T S Jayram, and Iordanis Kerenidis. Exponential separation of quantum and classical One-Way communication complexity. *SIAM J. Comput.*, 38(1):366–384, January 2008.

Paul Barham, Aakanksha Chowdhery, Jeff Dean, Sanjay Ghemawat, Steven Hand, Dan Hurt, Michael Isard, Hyeontaek Lim, Ruoming Pang, Sudip Roy, Brennan Saeta, Parker Schuh, Ryan Sepassi, Laurent El Shafey, Chandramohan A Thekkath, and Yonghui Wu. Pathways: Asynchronous distributed dataflow for ML. March 2022.

C H Bennett, G Brassard, C Crépeau, R Jozsa, A Peres, and W K Wootters. Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels. *Phys. Rev. Lett.*, 70(13):1895–1899, March 1993.

Charles H Bennett, Gilles Brassard, Sandu Popescu, Benjamin Schumacher, John A Smolin, and William K Wootters. Purification of noisy entanglement and faithful teleportation via noisy channels. November 1995.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D Manning, Suvir

Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W Thomas, Florian Tramèr, Rose E Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. August 2021.

Fernando G S Brandão, Amir Kalev, Tongyang Li, Cedric Yen-Yu Lin, Krysta M Svore, and Xiaodi Wu. Quantum SDP solvers: Large speed-ups, optimality, and applications to quantum learning. October 2017.

Gilles Brassard. Quantum communication complexity (a survey). January 2001.

Adam R Brown and Leonard Susskind. The second law of quantum complexity. January 2017.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are Few-Shot learners. May 2020.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. May 2014.

Harry Buhrman, Richard Cleve, Serge Massar, and Ronald de Wolf. Non-locality and communication complexity. *arXiv [quant-ph]*, July 2009.

M Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, and Patrick J Coles. Variational quantum algorithms. December 2020.

Shantanav Chakraborty, András Gilyén, and Stacey Jeffery. The power of block-encoded matrix powers: improved regression techniques via faster hamiltonian simulation. April 2018.

A Chi-Chih Yao. Quantum circuit complexity. In *Proceedings of 1993 IEEE 34th Annual Foundations of Computer Science*, pp. 352–361, November 1993.

Andrew M Childs and Wim van Dam. Quantum algorithms for algebraic problems. December 2008.

Nadav Cohen and Amnon Shashua. Inductive bias of deep convolutional networks through pooling geometry. May 2016.

Nadav Cohen, Or Sharir, and Amnon Shashua. On the expressive power of deep learning: A tensor analysis. September 2015.

Gavin E Crooks. Gradients of parameterized quantum gates using the parameter-shift rule and gate decomposition. May 2019.

Danny Driess, Fei Xia, Mehdi S M Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: An embodied multimodal language model. March 2023.

Edward Farhi and Hartmut Neven. Classification with quantum neural networks on near term processors. February 2018.

Richard P Feynman. Simulating physics with computers. *Int. J. Theor. Phys.*, 21(6):467–488, June 1982.

András Gilyén, Yuan Su, Guang Hao Low, and Nathan Wiebe. Quantum singular value transformation and beyond: exponential improvements for quantum matrix arithmetics. June 2018.

Vittorio Giovannetti, Seth Lloyd, and Lorenzo Maccone. Quantum random access memory. *Phys. Rev. Lett.*, 100(16): 160501, April 2008.

Lukas Gonon and Antoine Jacquier. Universal approximation theorem and error bounds for quantum neural networks and quantum reservoirs. July 2023.

Google Quantum AI. Suppressing quantum errors by scaling a surface code logical qubit. *Nature*, 614(7949):676–681, February 2023.

Goren Gordon and Gustavo Rigolin. Generalized teleportation protocol. November 2005.

Robert Gower, Donald Goldfarb, and Peter Richtarik. Stochastic block BFGS: Squeezing more curvature out of data. In Maria Florina Balcan and Kilian Q Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1869–1878, New York, New York, USA, 2016. PMLR.

Aram W Harrow and John C Napp. Low-Depth gradient measurements can improve convergence in variational hybrid Quantum-Classical algorithms. *Phys. Rev. Lett.*, 126(14):140502, April 2021.

Aram W Harrow, Avinatan Hassidim, and Seth Lloyd. Quantum algorithm for linear systems of equations. *Phys. Rev. Lett.*, 103(15):150502, October 2009.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W Rae, Oriol Vinyals, and Laurent Sifre. Training Compute-Optimal large language models. March 2022.

Alexander Semenovich Holevo. Bounds for the quantity of information transmitted by a quantum communication channel. *Problemy Peredachi Informatsii*, 1973.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. January 2018.

Hsin-Yuan Huang, Michael Broughton, Masoud Mohseni, Ryan Babbush, Sergio Boixo, Hartmut Neven, and Jarrod R McClean. Power of data in quantum machine learning. *Nat. Commun.*, 12(1):2631, May 2021.

Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Mia Xu Chen, Dehao Chen, Hyoukjoong Lee, Jiquan Ngiam, Quoc V Le, Yonghui Wu, and Zhifeng Chen. GPipe: Efficient training of giant neural networks using pipeline parallelism. November 2018.

William J Huggins and Jarrod R McClean. Accelerating quantum algorithms with precomputation. May 2023.

Rahul Jain, Jaikumar Radhakrishnan, and Pranab Sen. The quantum communication complexity of the pointer chasing problem: The bit version. In *FST TCS 2002: Foundations of Software Technology and Theoretical Computer Science*, Lecture notes in computer science, pp. 218–229. Springer Berlin Heidelberg, Berlin, Heidelberg, 2002.

Zhengfeng Ji, Yi-Kai Liu, and Fang Song. Pseudorandom quantum states. In *Lecture Notes in Computer Science*, Lecture notes in computer science, pp. 126–152. Springer International Publishing, Cham, 2018.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NeurIPS*, 2013.

Norman P Jouppi, George Kurian, Sheng Li, Peter Ma, Rahul Nagarajan, Lifeng Nai, Nishant Patil, Suvinay Subramanian, Andy Swing, Brian Towles, Cliff Young, Xiang Zhou, Zongwei Zhou, and David Patterson. TPU v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings. April 2023.

J Kaplan, S McCandlish, T Henighan, T B Brown, and others. Scaling laws for neural language models. *arXiv preprint arXiv*, 2020.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. June 2020.

V Krutyanskiy, M Galli, V Krcmarsky, S Baier, D A Fioretto, Y Pu, A Mazloom, P Sekatski, M Canteri, M Teller, J Schupp, J Bate, M Meraner, N Sangouard, B P Lanyon, and T E Northup. Entanglement of trapped-ion qubits separated by 230 meters. August 2022.

Eyal Kushilevitz and Noam Nisan. *Communication Complexity*. Cambridge University Press, Cambridge, England, April 2011.

Nikolai Lauk, Neil Sinclair, Shabir Barzanjeh, Jacob P Covey, Mark Saffman, Maria Spiropulu, and Christoph Simon. Perspectives on quantum transduction. *Quantum Sci. Technol.*, 5(2):020501, March 2020.

Nicolas Le Roux, Mark Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. February 2012.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. FNet: Mixing tokens with fourier transforms. May 2021.

Yoav Levine, Or Sharir, Alon Ziv, and Amnon Shashua. On the Long-Term memory of deep recurrent networks. October 2017.

Yoav Levine, Noam Wies, Or Sharir, Hofit Bata, and Amnon Shashua. The Depth-to-Width interplay in Self-Attention. June 2020.

Bo Li, Yuan Cao, Yu-Huai Li, Wen-Qi Cai, Wei-Yue Liu, Ji-Gang Ren, Sheng-Kai Liao, Hui-Nan Wu, Shuang-Lin Li, Li Li, Nai-Le Liu, Chao-Yang Lu, Juan Yin, Yu-Ao Chen, Cheng-Zhi Peng, and Jian-Wei Pan. Quantum state transfer over 1200 km assisted by prior distributed entanglement. *Phys. Rev. Lett.*, 128(17):170501, April 2022.

S Lloyd. Universal quantum simulators. *Science*, 273(5278):1073–1078, August 1996.

Seth Lloyd, Masoud Mohseni, and Patrick Rebentrost. Quantum principal component analysis. *Nat. Phys.*, 10(9): 631–633, July 2014.

Guang Hao Low and Isaac L Chuang. Optimal hamiltonian simulation by quantum signal processing. *Phys. Rev. Lett.*, 118(1):010501, January 2017.

P Magnard, S Storz, P Kurpiers, J Schär, F Marxer, J Lütolf, T Walter, J-C Besse, M Gabureac, K Reuer, A Akin, B Royer, A Blais, and A Wallraff. Microwave quantum link between superconducting circuits housed in spatially separated cryogenic systems. *Phys. Rev. Lett.*, 125(26):260502, December 2020.

Vitaly Maiorov and Allan Pinkus. Lower bounds for approximation by MLP neural networks. *Neurocomputing*, 25 (1):81–91, April 1999.

John M Martyn, Zane M Rossi, Andrew K Tan, and Isaac L Chuang. A grand unification of quantum algorithms. May 2021.

Jarrod R McClean, Jonathan Romero, Ryan Babbush, and Alán Aspuru-Guzik. The theory of variational hybrid quantum-classical algorithms. September 2015.

Hagay Michaeli, Tomer Michaeli, and Daniel Soudry. Alias-Free convnets: Fractional shift invariance via polynomial activations. March 2023.

Boris Mityagin. The zero set of a real analytic function. December 2015.

Ashley Montanaro and Sam Pallister. Quantum algorithms and the finite element method. December 2015.

Philipp Moritz, Robert Nishihara, and Michael Jordan. A Linearly-Convergent stochastic L-BFGS algorithm. In Arthur Gretton and Christian C Robert (eds.), *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pp. 249–258, Cadiz, Spain, 2016. PMLR.

Danial Motlagh and Nathan Wiebe. Generalized quantum signal processing. August 2023.

William J Munro, Koji Azuma, Kiyoshi Tamaki, and Kae Nemoto. Inside quantum repeaters. *IEEE J. Sel. Top. Quantum Electron.*, 21(3):78–90, May 2015.

Deepak Narayanan, Aaron Harlap, Amar Phanishayee, Vivek Seshadri, Nikhil R Devanur, Gregory R Ganger, Phillip B Gibbons, and Matei Zaharia. PipeDream: generalized pipeline parallelism for DNN training. In *Proceedings of the 27th ACM Symposium on Operating Systems Principles*, SOSP '19, pp. 1–15, New York, NY, USA, October 2019. Association for Computing Machinery.

Ashwin Nayak and Felix Wu. The quantum query complexity of approximating the median and related statistics. April 1998.

Michael A Nielsen and Isaac L Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press, December 2010.

Brad G Osgood. *Lectures on the Fourier Transform and Its Applications (Pure and Applied Undergraduate Texts) (Pure and Applied Undergraduate Texts, 33)*. American Mathematical Society, January 2019.

Adrián Pérez-Salinas, Alba Cervera-Lierta, Elies Gil-Fuster, and José I Latorre. Data re-uploading for a universal quantum classifier. July 2019.

M Pompili, S L N Hermans, S Baier, H K C Beukers, P C Humphreys, R N Schouten, R F L Vermeulen, M J Tiggelman, L Dos Santos Martins, B Dirkse, S Wehner, and R Hanson. Realization of a multinode quantum network of remote solid-state qubits. *Science*, 372(6539):259–264, April 2021.

Stephen J Ponzio, Jaikumar Radhakrishnan, and S Venkatesh. The communication complexity of pointer chasing. *J. Comput. System Sci.*, 62(2):323–355, March 2001.

Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Anselm Levskaya, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. Efficiently scaling transformer inference. November 2022.

Anup Rao and Amir Yehudayoff. *Communication Complexity and Applications*. Cambridge University Press, January 2020.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pp. 3505–3506, New York, NY, USA, August 2020. Association for Computing Machinery.

Arthur G Rattew and Patrick Rebentrost. Non-Linear transformations of quantum amplitudes: Exponential improvement, generalization, and applications. September 2023.

Ran Raz. Exponential separation of quantum and classical communication complexity. In *Proceedings of the thirty-first annual ACM symposium on Theory of Computing*, STOC '99, pp. 358–367, New York, NY, USA, May 1999. Association for Computing Machinery.

Alexander Razborov. Quantum communication complexity of symmetric predicates. April 2002.

Tim Roughgarden. Communication complexity (for algorithm designers). September 2015.

Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer. The effect of data encoding on the expressive power of variational quantum machine learning models. August 2020.

P W Shor. Algorithms for quantum computation: discrete logarithms and factoring. In *Proceedings 35th Annual Symposium on Foundations of Computer Science*. IEEE Comput. Soc. Press, 1994.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models. July 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł Ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008. Curran Associates, Inc., 2017.

A J Walker. New fast method for generating discrete random numbers with arbitrary frequency distributions. *Electron. Lett.*, 10(8):127–128, April 1974.

Changqing Wang, Ivan Gonin, Anna Grassellino, Sergey Kazakov, Alexander Romanenko, Vyacheslav P Yakovlev, and Silvia Zorzetti. High-efficiency microwave-optical quantum transduction based on a cavity electro-optic superconducting system with long coherence time. *npj Quantum Information*, 8(1):1–10, December 2022.

Yuanzhong Xu, Hyoukjoong Lee, Dehao Chen, Blake Hechtman, Yanping Huang, Rahul Joshi, Maxim Krikun, Dmitry Lepikhin, Andy Ly, Marcello Maggioni, Ruoming Pang, Noam Shazeer, Shibo Wang, Tao Wang, Yonghui Wu, and Zhifeng Chen. GSPMD: General and scalable parallelization for ML computation graphs. May 2021.

Andrew Chi-Chih Yao. Some complexity questions related to distributive computing(preliminary report). In *Proceedings of the eleventh annual ACM symposium on Theory of computing*, STOC '79, pp. 209–213, New York, NY, USA, April 1979. Association for Computing Machinery.